OPINION ARTICLE

REVISED Developing WHO guidelines: Time to formally include evidence from mathematical modelling studies [version 2; referees: 2 approved]

Matthias Egger [1,2], Leigh Johnson[2], Christian Althaus[1], Anna Schöni[1], Georgia Salanti[1], Nicola Low [1], Susan L. Norris [3]

[1]Institute of Social and Preventive Medicine (ISPM), University of Bern, Bern, 3012, Switzerland
[2]Centre for Infectious Disease Epidemiology and Research (CIDER), University of Cape Town, Cape Town, 7925, South Africa
[3]World Health Organization, Geneva, Switzerland

v2 **First published:** 29 Aug 2017, **6**:1584 (doi: 10.12688/f1000research.12367.1)
Latest published: 26 Feb 2018, **6**:1584 (doi: 10.12688/f1000research.12367.2)

Abstract

In recent years, the number of mathematical modelling studies has increased steeply. Many of the questions addressed in these studies are relevant to the development of World Health Organization (WHO) guidelines, but modelling studies are rarely formally included as part of the body of evidence. An expert consultation hosted by WHO, a survey of modellers and users of modelling studies, and literature reviews informed the development of recommendations on when and how to incorporate the results of modelling studies into WHO guidelines. In this article, we argue that modelling studies should routinely be considered in the process of developing WHO guidelines, but particularly in the evaluation of public health programmes, long-term effectiveness or comparative effectiveness. There should be a systematic and transparent approach to identifying relevant published models, and to commissioning new models. We believe that the inclusion of evidence from modelling studies into the Grading of Recommendations Assessment, Development and Evaluation (GRADE) process is possible and desirable, with relatively few adaptations. No single “one-size-fits-all” approach is appropriate to assess the quality of modelling studies. The concept of the ‘credibility’ of the model, which takes the conceptualization of the problem, model structure, input data, different dimensions of uncertainty, as well as transparency and validation into account, is more appropriate than ‘risk of bias’.

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees 1	2
REVISED version 2 published 26 Feb 2018		✓ report
version 1 published 29 Aug 2017	✓ report	? report

1 **Wilma A. Stolk**, University Medical Center Rotterdam, Netherlands

2 **Joseph W Hogan**, Brown University School of Public Health, USA

Discuss this article

Comments (0)

Corresponding author: Susan L. Norris (norriss@who.int)

Author roles: Egger M: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Writing – Original Draft Preparation; **Johnson L**: Conceptualization, Funding Acquisition, Writing – Review & Editing; **Althaus C**: Conceptualization, Funding Acquisition, Investigation, Methodology, Resources, Software, Supervision, Writing – Review & Editing; **Schöni A**: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Visualization, Writing – Review & Editing; **Salanti G**: Conceptualization, Funding Acquisition, Writing – Review & Editing; **Low N**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Resources, Software, Supervision, Visualization, Writing – Review & Editing; **Norris SL**: Conceptualization, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: Susan L. Norris is a member of the GRADE working group. No other competing interests were disclosed.

How to cite this article: Egger M, Johnson L, Althaus C *et al.* **Developing WHO guidelines: Time to formally include evidence from mathematical modelling studies [version 2; referees: 2 approved]** *F1000Research* 2018, **6**:1584 (doi: 10.12688/f1000research.12367.2)

Grant information: The work reported in this article and the expert consultation meeting in Geneva, Switzerland, were funded by the UNICEF/UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases (WHO/TDR).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

First published: 29 Aug 2017, **6**:1584 (doi: 10.12688/f1000research.12367.1)

REVISED **Amendments from Version 1**

We have clarified and elaborated upon the distinctions between mathematical and statistical modelling and between a mathematical model and a mathematical modelling study. We use a broad definition of mathematical models which encompasses both descriptive and predictive aspects. Statistical modelling, on the other hand, typically characterizes sources of variation and associations between variables in observed populations of interest. We also elaborate on the GRADE domain of risk of bias as part of the assessment of certainty of a body of evidence for important and critical outcomes. We feel that the concept of risk of bias is too narrow in the context of mathematical modelling studies and prefer to use "credibility" which encompasses not only by risk of bias of the input data, but also conceptualization of the problem, model structure, other dimensions of uncertainty, transparency, and validation.

See referee reports

Introduction

Mathematical models have a long history in public health[1]. In 1760, Daniel Bernoulli developed a model of smallpox transmission and control. William Hamer published a measles transmission model in 1906 and Ronald Ross a model of malaria transmission in 1908. In recent years, the number of publications related to mathematical modelling has increased steeply. Today, mathematical modelling studies are not restricted to infectious diseases but address a wide range of questions.

The World Health Organization (WHO) provides recommendations on many public health, health system and clinical topics. WHO guidelines are developed using processes and methods that ensure the publication of high-quality recommendations, as outlined in the *WHO Handbook for Guideline Development*[2]. WHO uses the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach to rate the certainty of a body of evidence and to produce information that is used by guideline panels to formulate recommendations, based on the balance of benefits and harms and other considerations[3].

Many of the questions addressed in mathematical modelling studies are relevant to the development of guidelines. Increasingly, WHO and other guideline developers need to decide whether and how the results of mathematical modelling studies should be included in the evidence base used to develop recommendations. We reviewed the 185 WHO guidelines that were approved by the Guidelines Review Committee from 2007 to 2015: 42 (23%) referred to mathematical modelling studies. However, these studies were rarely formally assessed as part of the body of evidence, and quality criteria for modelling studies were often lacking. A major barrier to the incorporation of evidence from mathematical modelling studies into guidelines is the perceived complexity of the methods used to construct and analyse these studies. At present, there are no widely agreed methods for, or approaches to, the evaluation of the results of mathematical modelling studies, and to their integration with primary data to inform guidelines and recommendations. In April 2016 WHO organized a workshop in Geneva, Switzerland to discuss when and how to incorporate the results of modelling studies into WHO guidelines (see *Acknowledgements* for names of participants). Specifically, the workshop participants discussed the following three questions::

(1) When is it appropriate to consider modelling studies as part of the evidence that supports a guideline?

(2) How should the quality and risk of bias in mathematical modelling studies be assessed?

(3) How can the GRADE approach be adapted to assess the certainty of a body of evidence that includes the results of modelling and to formulate recommendations?

A detailed workshop report is available from WHO[4].

The role of modelling in economic evaluation is well recognised in guideline development and at WHO, and was therefore excluded from discussions. At the workshop, we considered the results of a survey of experts (see Box 1) and a rapid literature review (see below). In this paper, which reflects the opinions of

Box 1. Web-based expert survey on the role of mathematical modelling in guideline development

The survey was conducted between March 17 and April 4, 2016. It consisted of 10 questions: four on the characteristics of the respondents, three on the role of mathematical models in guideline development, two questions on quality criteria for mathematical models and one on the challenges in using mathematical modelling in guideline development (see Figure S1). Using snowball sampling, mathematical modellers, epidemiologists, guideline developers and other experts were invited to participate in the survey. A total of 151 individuals from 28 countries and 87 different institutions responded. About half of respondents were modellers, and the other half users of the results from modelling studies. The majority of respondents (58%) had been part of a guideline development group in the past.

Ninety-five percent of respondents answered yes to the question "Should mathematical modelling inform guidance for public health interventions?" and 60% indicated that findings of mathematical modelling studies can sometimes provide the same level of evidence as those of empirical research studies. When asked to list situations in which mathematical modelling could be particularly useful for the development of guidelines, the absence of empirical data on the effectiveness, cost-effectiveness and impact of an intervention, and on the comparative effectiveness of different interventions was most frequently mentioned. We also asked about situations where mathematical modelling studies should not be used or have been inappropriately used in the development of guidelines. Respondents reported that modelling should not be used "to cover up" for the lack of evidence from empirical research, and due emphasis should be given to the uncertainty of model predictions. When asked about the five most important criteria for the quality of reporting of modelling studies, respondents mentioned that the model structure should be clearly described and justified, the important sources of uncertainty reported, and model validity addressed. Assumptions should be clearly stated, justified and discussed and the sources of parameter estimates described. Finally, respondents identified the interpretation of results from modelling studies, the evaluation of their quality and the communication of uncertainty as major challenges in using mathematical modelling in guideline development. These challenges would be best addressed by including at least one modelling expert in guideline development groups.

the authors but not necessarily that of all workshop participants, we first define models and modelling studies. We then address the three questions outlined above and conclude with some recommendations on the use of evidence from modelling studies in guidelines development.

What is a mathematical modelling study?

Using a common terminology across different disciplines, for example infectious disease modelling and modelling in chronic disease, will facilitate the assessment, evaluation and comparison of mathematical modelling studies. A broad definition of a *mathematical model* is a "mathematical framework representing variables and their interrelationships to describe observed phenomena or predict future events"[5]. We make a distinction between a mathematical model and *mathematical modelling studies*, which we define as studies that address defined research questions using mathematical modelling. Mathematical modelling studies typically address complex situations and tend to rely more heavily on assumptions about underlying mathematical structure than on individual-level data. Examples include investigating the potential of HIV testing with immediate antiretroviral therapy to reduce HIV transmission[6], or the likely impact of different screening practices on the incidence of cervical cancer[7].

Statistical modelling is typically concerned with characterizing sources of variation and associations between variables in observed individual-level data drawn from a target population of interest and tends to address questions of a narrower scope than mathematical models. Both statistical and mathematical models can be used to predict future outcomes and to compare different policies. The results from statistical analyses of empirical data often inform mathematical models. Mathematical modelling studies also increasingly integrate statistical models to relate the model output to data.

Workshop participants discussed whether it might be helpful for guideline groups to classify mathematical models in terms of their scope (for example descriptive versus predictive), or technical approach (for example static versus dynamic)[8]. Discussants argued that a good understanding of what information models can provide and what level of confidence can be placed in that information was more important than a detailed taxonomy of models[4].

Role of mathematical modelling studies in guideline development

Mathematical models typically address questions that cannot easily be answered with randomized controlled trials (RCTs) or observational studies. Table 1 lists specific situations and examples where the results of mathematical modelling are particularly relevant to guideline development, based on the survey, published examples and the Geneva workshop. Mathematical modelling can overcome some of the limitations of results obtained from the carefully controlled settings in which RCTs are typically conducted. First, the main trial results provide an average effect estimate that applies to a specific intervention and study population. Mathematical modelling studies can be used to extrapolate from the results of RCTs to different target groups and

Table 1. Situations in which mathematical modelling studies may be useful for guideline development.

Situation	Examples of relevant mathematical modelling studies
The long-term effectiveness or cost-effectiveness of an intervention is unclear.	Life time effect on decompensated cirrhosis of obeticholic acid as second-line treatment in primary biliary cholangitis[9]. Outcomes and costs over 10 years of donepezil treatment in mild to moderately severe Alzheimer's Disease[10]. Long-term clinical outcomes, costs and cost-effectiveness of interventions in diabetes mellitus (types 1 and 2)[11].
The outcomes of an intervention in real world, routine care settings are unclear.	Outcomes of medical management of asymptomatic patients with carotid artery stenosis who were excluded from clinical trials[12]. Effects on blood pressure and cardiovascular risk of variations in patients' adherence to prescribed antihypertensive drugs[13].
The comparative (relative) effectiveness of different interventions overall or in subgroups of patients is unclear.	Comparative effectiveness of different statins and statin doses in patient groups with varying baseline cardiovascular risk[14]. Relative effect of different strategies of incorporating bevacizumab into platinum-based treatment on survival in ovarian cancer[15]. Relative real-world drug effectiveness of disease modifying anti-rheumatic drugs (DMARDs)[16].
The overall effects of an intervention at the population level, including direct and indirect effects, are unknown.	Effects of different vaccination strategies with serogroup C meningococcal conjugate vaccines on meningococcal carriage and disease[17]. Public health impact of vaccinating boys and men with a quadrivalent HPV vaccine[18]. Impact of expanding access to antiretroviral therapy ("treatment as prevention") on new HIV infections[19].
The population burden of a disease or condition is unknown.	Estimate of the global burden of latent tuberculosis infection[20]. Burden of healthcare-associated infections on European population health[21]. Global variation in stroke burden and mortality[22].

Source: WHO expert survey and consultation.

settings, to long term outcomes, and to bridge the gap between efficacy and (long-term) effectiveness[23]. Second, interventions to prevent and control infectious diseases have non-linear effects. RCTs that address short term effects at the individual level might not be suitable for estimating the longer term effects of introducing an intervention, say a vaccine, in a whole population if indirect herd effects influence the incidence of infection and hence the impact of the intervention[24,25]. Third, rapid guidance is often needed early in outbreaks or public health emergencies when relevant interventions for prevention or management might simply not have been evaluated. The results of mathematical modelling studies can be used to draft emergency guidelines or to assess the epidemic potential of new outbreaks[26].

The findings of mathematical modelling studies are only as good as the data and assumptions that inform them. Guideline recommendations should therefore not be based on the outputs of models when uncertainty in the empirical data has not been appropriately quantified, when the model makes implausible assumptions or has not been validated adequately, or when the model predictions vary widely over a plausible range of parameter estimates.

Assessing the quality of a mathematical modelling study: Rapid review

We performed a rapid review of the methodological literature to identify criteria that are proposed to assess the "quality" of mathematical modelling studies (see Table S1 for the detailed search strategy). Specifically, we aimed to identify criteria proposed to assess the quality of single mathematical modelling studies, including best practice standards or criteria for assessing risk of bias or reporting quality and criteria proposed to assess the quality of a body of evidence from mathematical modelling studies. We were also interested in identifying checklists or other instruments developed to assess the quality of mathematical modelling studies.

We identified 20 relevant articles (see Figure 1 for a flow chart of the identification of eligible articles)[25,27–44]. Most gave recommendations for good modelling practice and were compiled by a task force in a consensus process or based on a systematic or narrative review of the literature. The widely cited 2003 paper by Weinstein and colleagues organized 28 recommendations under the headings "structure", "data", and "validation"[31]. A questionnaire or checklist was not included. A subsequent series of seven articles[25,38–42,44] by the joint International Society for Pharmacoeconomics and Outcomes Research (ISPOR) and Society for Medical Decision Making (SMDM) task force elaborated upon these recommendations, providing detailed advice on conceptualizing the model, state transition models, discrete event simulations, dynamic transmission models, parameter estimation and uncertainty, and transparency and validation. The 79 recommendations are summarized in the first article of the series[44].

We identified four articles[32,34,37,43] that present comprehensive frameworks of good modelling practice, with detailed justifications of the items covered and attributes of good practice. They include signalling or helper questions to facilitate the critical appraisal of

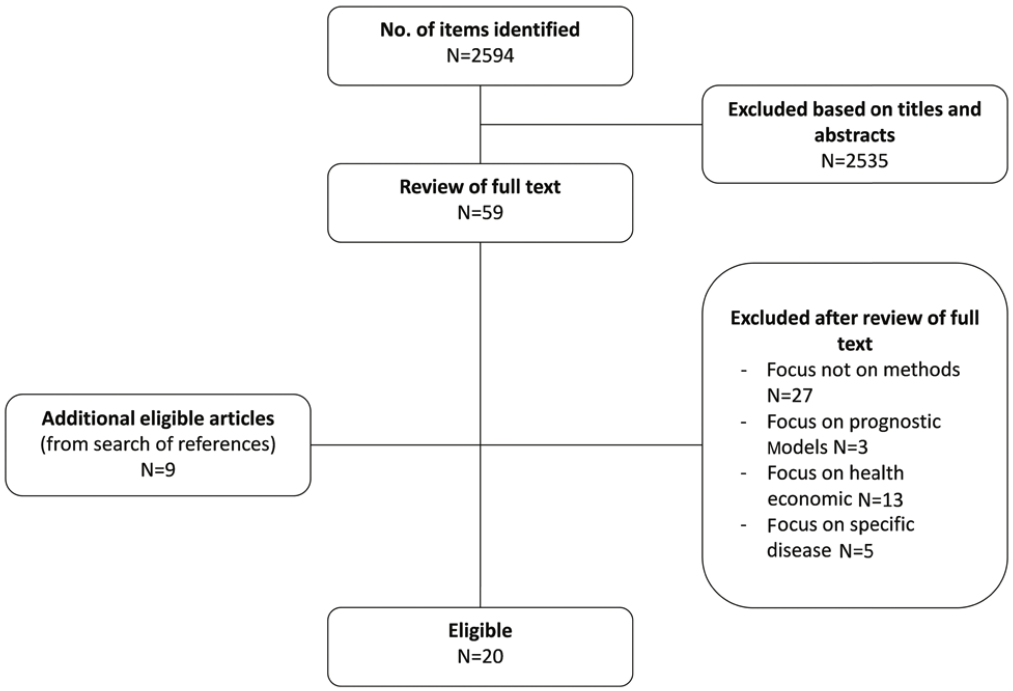

Figure 1. Rapid review of literature on good practice in mathematical modelling: flow of identification of eligible studies.

published modelling studies: the number of questions ranges from 38 in Caro *et al.*[32] to 66 questions in Bennett and Manuel[37]. The four frameworks cover similar territory, including items related to the problem concept, model structure, data sources and synthesis of the evidence, model uncertainty, consistency, transparency and validation (Table 2). Two of the frameworks include sponsorship and conflicts of interest[32,37].

In a qualitative study Chilcot *et al.*[27] performed in-depth interviews with 12 modellers from academic and commercial sectors, and model credibility emerged as the central concern of decision-makers using models. Respondents agreed that developing an understanding of the clinical situation or disease process being investigated is paramount in ensuring model credibility, highlighting the importance of clinical input during the model development process[27].

Model comparisons and modelling consortia

Published mathematical models addressing the same issue may reach contrasting conclusions. In this situation, careful comparison of the models may lead to a deeper understanding of the factors that drive outputs and conclusions. Ideally, the different modelling groups come together to explore the importance of differences in the type and structure of their models, and of the data used to parameterize them[19,45,46]. For example, several groups of modellers have investigated the impact of expanding access to antiretroviral therapy (ART) on new HIV infections. The HIV Modelling Consortium compared the predictions of several mathematical models simulating the same ART intervention programs to determine the extent to which models agree on the epidemiological impact of expanded ART[19]. The consortium concluded that although models vary substantially in structure, complexity, and parameter choices, all suggested that ART, at high levels of access

Table 2. Items covered by four published frameworks developed to assess good modelling practice.

Philips 2006[34]	Bennett 2012[37]	Caro 2014[32]	Peñaloza Ramos 2015[43]
Structure Decision problem/objective Scope/perspective Rationale for structure Structural assumptions Strategies/comparators Model type Time horizon Disease states/pathways Cycle length **Data** Data identification Pre-model data analysis Baseline data Treatment effects Utilities Data incorporation Assessment of uncertainty Methodological Structural Heterogeneity Parameter **Consistency** Internal consistency External consistency	**Structure** Decision problem/objective Scope/perspective Rationale for structure Structural assumptions Strategies/comparators Model type Time horizon Disease states/pathways Cycle length Parsimony **Data** Data identification Data modelling Baseline data Treatment effects Risk factors Data incorporation Assessment of uncertainty Methodological Structural Heterogeneity Parameter **Consistency** Internal consistency External consistency **Validity** Output plausibility Predictive validity **Computer implementation** **Transparency** **Sponsorship**	**RELEVANCE** Population Interventions Outcomes Context **CREDIBILITY** **Validation** External validation Internal verification Face validity **Design** Problem concept Model concept and structure **Data** Process of obtaining and values of inputs **Analysis** Adequacy Uncertainly **Reporting** Adequacy **Interpretation** Balance **Conflict of interest** Potential conflicts and steps taken to address them	**Problem concept** Decision problem Analytical perspective Target population Health outcomes Comparators Time horizon **Model concept** Choice of model type Model structure **Synthesis of evidence** Data sources Utilities Cycle length and half-cycle correction Resources/costs Patient heterogeneity Parameter precision **Model uncertainty** Analyses of uncertainty related to the decision problem Parameter estimation Structural uncertainty Other analyses of uncertainty **Model transparency and validation** Transparency Validation Face validity Internal validity Cross-validation External validity Predictive validity

and with high adherence, has the potential to substantially reduce new HIV infections in the population[19]. There was broad agreement regarding the short-term epidemiologic impact of ART scale-up, but more variation in longer-term projections and in the efficiency with which treatment can reduce new infections. The impact of ART on HIV incidence long-term is expected to be lower if models: (i) allow for heterogeneity in sexual risk behaviour; (ii) are age-structured; (iii) estimate a low proportion of HIV transmission from individuals not on ART with advanced disease (at low CD4 counts); (iv) are compared to what would be expected in the presence of HIV counselling and testing (compared to no counselling and testing); (v) assume relatively high infectiousness on ART; and (vi) consider drug resistance[19,47,48].

Assessing mathematical modelling studies using the GRADE approach

GRADE was conceived with the intention of creating a uniform system to assess a body of evidence to support guideline development in response to a confusing array of different systems in use at that time[49]. It has since been adopted by over 90 organisations, including WHO. GRADE addresses clinical management questions, including the impact of therapies and diagnostic strategies, diagnostic accuracy questions (i.e., the accuracy of a single diagnostic or screening test), the (cost-) effectiveness and safety of public health interventions, and questions about prognosis.

The GRADE approach encompasses two main considerations: the *degree of certainty* in the evidence used to support a decision and *the strength of the recommendation*. The degree of certainty, i.e., the confidence in or quality of a body of evidence, is rated as "high", "moderate", "low", or "very low" based on an assessment of five dimensions: study limitations (risk of bias), imprecision, inconsistency, indirectness, and publication bias. The initial assessment is based on the study design: RCTs start as high certainty and observational studies as low certainty. Based on the assessments of the five dimensions, RCTs may be down-rated and observational studies up- or down-rated. Judgment is required when assessing the certainty of the evidence, taking into account the number of studies of higher and lower quality and the relative importance of the different dimensions in a given context. The second consideration is the strength of the recommendation, which can be "strong" or "conditional", for or against an intervention or test, based on the balance of benefits and harms, certainty of the evidence, the relative values of persons affected by the intervention, resource considerations, acceptability and feasibility, among others[50].

We believe that evidence from mathematical modelling studies could be assessed within the GRADE framework and included in the guideline development process. Specifically, guideline groups might include mathematical modelling studies as an additional study category, in addition to the categories of RCTs and observational studies currently defined in GRADE. The dimensions of indirectness, inconsistency, imprecision and publication bias are applicable to mathematical modelling studies, but criteria may need to be adapted. The concept of bias relates to results or inferences from empirical studies, including RCTs and observational studies[51,52] and is too narrow in the context of assessing mathematical modelling studies[53]. "Credibility", a term used by ISPOR[54], may therefore be more appropriate for modelling studies than "risk of bias". The assessment of the credibility of a model is informed by a comprehensive quality framework and should cover the conceptualization of the problem, model structure, input data and their risk of bias, different dimensions of uncertainty, as well as transparency and validation (Table 2). The framework should be tailored to each set of modelling studies by adding or omitting questions and developing review-specific guidance on how to assess each criterion. The certainty of the body of evidence from modelling studies can then be classified as high, moderate, low, or very low. In the evidence-to-decision framework a distinction should be made between *observed outcomes* from empirical studies and *modelled outcomes* from modelling studies (see the Meeting Report[4] for an example).

Conclusions and recommendations

Based on the discussions and presentations at the workshop in Geneva, the survey and rapid systematic review, we believe a number of conclusions can be formulated.

When is it appropriate to consider modelling studies as part of the evidence that supports a guideline?

1. The use of modelling studies should routinely be considered in the process of developing WHO guidelines. Findings of mathematical modelling studies can provide important evidence that may be highly relevant. Evidence from modelling studies should be considered specifically in the absence of empirical data directly addressing the question of interest, when modelling based on appropriate indirect evidence may be indicated. Examples for such situations include the evaluation of long-term effectiveness, and the impact of one or several interventions (comparative effectiveness), for example in the context of public health programmes where RCTs are rarely available.

2. Modelling may be more acceptable and more influential in situations where immediate action is called for, but little direct empirical evidence is available, and may arguably be more acceptable in public health than in clinical decision making. In these situations (for example, the HIV, Ebola, or Zika epidemics) funding is also likely to become available to support dedicated modelling studies.

3. The use of evidence from mathematical models should be carefully considered and there should be a systematic and transparent approach to identifying existing models that may be relevant, and to commissioning new models.

How should the credibility of mathematical modelling studies be assessed?

4. No single "one-size-fits-all" approach is appropriate to assess the quality of modelling studies. Existing frameworks and checklists may be adapted to a set of modelling studies by adding or omitting questions. In some situations, the approach will need to be developed *de novo*.

5. Additional expertise will typically be required in the systematic review groups or guideline development groups to appropriately assess the credibility of modelling studies and interpret their results.

6. The credibility of the models should not be evaluated only by modellers, and not only by modellers involved in the development of these models.

How can the GRADE approach be adapted to assess a body of evidence that includes the results of modelling and to formulate recommendations?

7. The inclusion of evidence from modelling studies into the GRADE process is possible and desirable, with relatively few adaptations. GRADE is simply rating the certainty of evidence to support a decision and any type of evidence can in principle be included.

8. The certainty of the evidence for modelling studies should be assessed and presented separately in summaries of the evidence (GRADE evidence profiles), and classified as high, moderate, low, or very low certainty.

9. The GRADE dimensions of certainty (imprecision, indirectness, inconsistency and publication bias) and the criteria defined for their assessment are also relevant to modelling studies.

10. For modelling studies, the concept of the 'credibility' of the model, which takes the structure of the model, input data, dimensions of uncertainty, as well as transparency and validation into account, is more appropriate than 'study limitations' or 'risk of bias'.

11. When summarizing the evidence, a distinction should be made between observed and modelled outcomes.

12. We propose that within the GRADE system, modelling studies start at low certainty. It should then be possible to increase or decrease the certainty of modelling studies based on a set of criteria. The development of these criteria was beyond the scope of this article; a GRADE working group is addressing this issue (http://www.gradeworkinggroup.org/).

We look forward to discussing these recommendations with experts and stakeholders and to developing exact procedures and criteria for the assessment of modelling studies and their inclusion in the GRADE process.

Competing interests

Susan L. Norris is a member of the GRADE working group. No other competing interests were disclosed.

Grant information

The work reported in this article and the expert consultation meeting in Geneva, Switzerland, were funded by the UNICEF/UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases (WHO/TDR).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We are grateful to Helen Ward and Tom Trikalinos who chaired the April 2016 expert meeting in Geneva and to all who participated in the meeting: Patrick Bossuyt, David Fisman, Gordon Guyatt, Tim Hallett, Mark Helfand, Rod Jackson, Veena Manja, Holger Schünemann, Julie Ann Simpson, Christopher Dye, Philippa Easterbrook, Nathan Ford, Daniel Hogan, and Gretchen Stevens. All authors of this article also participated.

Supplementary material

Table S1. Search strategy in MEDLINE from inception to January 2016 without language restrictions, combining terms for mathematical models with terms for quality assessment and health care decision-making.

Click here to access the data.

Figure S1. Questionnaire of the online survey on the use of mathematical modelling in guidelines for public health decision making.

Click here to access the data.

References

1. Smith DL, Battle KE, Hay SI, *et al.*: **Ross, macdonald, and a theory for the dynamics and control of mosquito-transmitted pathogens.** *PLoS Pathog.* 2012; **8**(4): e1002588.
PubMed Abstract | Publisher Full Text | Free Full Text
2. World Health Organization: **WHO Handbook for Guideline Development.** Geneva, 2014.
Reference Source
3. Guyatt GH, Oxman AD, Vist GE, *et al.*: **GRADE: An emerging consensus on rating quality of evidence and strength of recommendations.** *BMJ.* 2008; **336**(7650): 924–6.
PubMed Abstract | Publisher Full Text | Free Full Text
4. World Health Organization: **Meeting report**. Geneva. 2017.
Reference Source
5. Eykhoff P: **System identification; parameter and state estimation**. Chester: John Wiley & Sons Ltd., 1974.
Reference Source

6. Granich RM, Gilks CF, Dye C, *et al.*: **Universal voluntary HIV testing with immediate antiretroviral therapy as a strategy for elimination of HIV transmission: a mathematical model.** *Lancet.* 2009; **373**(9657): 48–57.
PubMed Abstract | Publisher Full Text

7. Canfell K, Barnabas R, Patnick J, *et al.*: **The predicted effect of changes in cervical screening practice in the UK: results from a modelling study.** *Br J Cancer.* 2004; **91**(3): 530–6.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Bolker BM: **Ecological models and data in R**. Princeton, NJ: Prineceton University Press. 2008.
Reference Source

9. Samur S, Klebanoff M, Banken R, *et al.*: **Long-term clinical impact and cost-effectiveness of obeticholic acid for the treatment of primary biliary cholangitis.** *Hepatology.* 2017; **65**(3): 920–928.
PubMed Abstract | Publisher Full Text

10. Getsios D, Blume S, Ishak KJ, *et al.*: **Cost effectiveness of donepezil in the treatment of mild to moderate Alzheimer's disease: a UK evaluation using discrete-event simulation.** *Pharmacoeconomics.* 2010; **28**(5): 411–27.
PubMed Abstract | Publisher Full Text

11. Palmer AJ, Roze S, Valentine WJ, *et al.*: **The CORE Diabetes Model: Projecting long-term clinical outcomes, costs and cost-effectiveness of interventions in diabetes mellitus (types 1 and 2) to support clinical and reimbursement decision-making.** *Curr Med Res Opin.* 2004; **20**(Suppl 1): S5–26.
PubMed Abstract | Publisher Full Text

12. Smolen HJ, Cohen DJ, Samsa GP, *et al.*: **Development, validation, and application of a microsimulation model to predict stroke and mortality in medically managed asymptomatic patients with significant carotid artery stenosis.** *Value Health.* 2007; **10**(6): 489–97.
PubMed Abstract | Publisher Full Text

13. Lowy A, Munk VC, Ong SH, *et al.*: **Effects on blood pressure and cardiovascular risk of variations in patients' adherence to prescribed antihypertensive drugs: role of duration of drug action.** *Int J Clin Pract.* 2011; **65**(1): 41–53.
PubMed Abstract | Publisher Full Text

14. Schuetz CA, van Herick A, Alperin P, *et al.*: **Comparing the effectiveness of rosuvastatin and atorvastatin in preventing cardiovascular outcomes: estimates using the Archimedes model.** *J Med Econ.* 2012; **15**(6): 1118–29.
PubMed Abstract | Publisher Full Text

15. Barnett JC, Alvarez Secord A, Cohn DE, *et al.*: **Cost effectiveness of alternative strategies for incorporating bevacizumab into the primary treatment of ovarian cancer.** *Cancer.* 2013; **119**(20): 3653–61.
PubMed Abstract | Publisher Full Text

16. Didden E, Ruffieux Y, Hummel N, *et al.*: **Prediction of Real-World Drug Effectiveness Pre-Launch: Case study in Rheumatoid Arthritis.** *Value Health.* 2017; submitted.

17. Trotter CL, Gay NJ, Edmunds WJ: **Dynamic models of meningococcal carriage, disease, and the impact of serogroup C conjugate vaccination.** *Am J Epidemiol.* 2005; **162**(1): 89–100.
PubMed Abstract | Publisher Full Text

18. Elbasha EH, Dasbach EJ: **Impact of vaccinating boys and men against HPV in the United States.** *Vaccine.* 2010; **28**(42): 6858–67.
PubMed Abstract | Publisher Full Text

19. Eaton JW, Johnson LF, Salomon JA, *et al.*: **HIV treatment as prevention: systematic comparison of mathematical models of the potential impact of antiretroviral therapy on HIV incidence in South Africa.** *PLoS Med.* 2012; **9**(7): e1001245.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Houben RM, Dodd PJ: **The Global Burden of Latent Tuberculosis Infection: A Re-estimation Using Mathematical Modelling.** *PLoS Med.* 2016; **13**(10): e1002152.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Cassini A, Plachouras D, Eckmanns T, *et al.*: **Burden of Six Healthcare-Associated Infections on European Population Health: Estimating Incidence-Based Disability-Adjusted Life Years through a Population Prevalence-Based Modelling Study.** *PLoS Med.* 2016; **13**(10): e1002150.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Johnston SC, Mendis S, Mathers CD: **Global variation in stroke burden and mortality: estimates from monitoring, surveillance, and modelling.** *Lancet Neurol.* 2009; **8**(4): 345–54.
PubMed Abstract | Publisher Full Text

23. Egger M, Moons KG, Fletcher C, *et al.*: **GetReal: from efficacy in clinical trials to relative effectiveness in the real world.** *Res Synth Methods.* 2016; **7**(3): 278–81.
PubMed Abstract | Publisher Full Text

24. Weinstein MC: **Recent developments in decision-analytic modelling for economic evaluation.** *Pharmacoeconomics.* 2006; **24**(11): 1043–53.
PubMed Abstract | Publisher Full Text

25. Pitman R, Fisman D, Zaric GS, *et al.*: **Dynamic transmission modeling: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--5.** *Value Health.* 2012; **15**(6): 828–34.
PubMed Abstract | Publisher Full Text

26. Camacho A, Kucharski AJ, Funk S, *et al.*: **Potential for large outbreaks of Ebola virus disease.** *Epidemics.* 2014; **9**: 70–8.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Chilcott J, Tappenden P, Rawdin A, *et al.*: **Avoiding and identifying errors in health technology assessment models: qualitative study and methodological review.** *Health Technol Assess.* 2010; **14**(25): iii–iv, ix–xii, 1–107.
PubMed Abstract | Publisher Full Text

28. Weinstein MC, Toy EL, Sandberg EA, *et al.*: **Modeling for health care and other policy decisions: uses, roles, and validity.** *Value Health.* 2001; **4**(5): 348–61.
PubMed Abstract | Publisher Full Text

29. Philips Z, Ginnelly L, Sculpher M, *et al.*: **Review of guidelines for good practice in decision-analytic modelling in health technology assessment.** *Health Technol Assess.* 2004; **8**(36): iii–iv, ix–xi, 1–158.
PubMed Abstract | Publisher Full Text

30. Goldhaber-Fiebert JD, Stout NK, Goldie SJ: **Empirically evaluating decision-analytic models.** *Value Health.* 2010; **13**(5): 667–74.
PubMed Abstract | Publisher Full Text

31. Weinstein MC, O'Brien B, Hornberger J, *et al.*: **Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices--Modeling Studies.** *Value Health.* 2003; **6**(1): 9–17.
PubMed Abstract | Publisher Full Text

32. Jaime Caro J, Eddy DM, Kan H, *et al.*: **Questionnaire to assess relevance and credibility of modeling studies for informing health care decision making: An ISPOR-AMCP-NPC good practice task force report.** *Value Health.* 2014; **17**(2): 174–82.
PubMed Abstract | Publisher Full Text

33. Bilcke J, Beutels P, Brisson M, *et al.*: **Accounting for Methodological, Structural, and Parameter Uncertainty in Decision-Analytic Models: A Practical Guide.** *Med Decis Mak.* 2011; **31**(4): 675–92.
PubMed Abstract | Publisher Full Text

34. Philips Z, Bojke L, Sculpher M, *et al.*: **Good practice guidelines for decision-analytic modelling in health technology assessment: a review and consolidation of quality assessment.** *Pharmacoeconomics.* 2006; **24**(4): 355–71.
PubMed Abstract | Publisher Full Text

35. Chilcott J, Brennan A, Booth A, *et al.*: **The role of modelling in prioritising and planning clinical trials.** *Health Technol Assess.* 2003; **7**(23): iii, 1–125.
PubMed Abstract | Publisher Full Text

36. Forsberg HH, Aronsson H, Keller C, *et al.*: **Managing health care decisions and improvement through simulation modeling.** *Qual Manag Health Care.* 2011; **20**(1): 15–29.
PubMed Abstract | Publisher Full Text

37. Bennett C, Manuel DG: **Reporting guidelines for modelling studies.** *BMC Med Res Methodol.* 2012; **12**: 168.
PubMed Abstract | Publisher Full Text | Free Full Text

38. Siebert U, Alagoz O, Bayoumi AM, *et al.*: **State-transition modeling: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--3.** *Value Health.* 2012; **15**(6): 812–20.
PubMed Abstract | Publisher Full Text

39. Briggs AH, Weinstein MC, Fenwick EA, *et al.*: **Model parameter estimation and uncertainty: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--6.** *Value Health.* 2012; **15**(6): 835–42.
PubMed Abstract | Publisher Full Text

40. Karnon J, Stahl J, Brennan A, *et al.*: **Modeling using Discrete Event Simulation: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force--4.** *Value Health.* 2012; **15**(6): 821–7.
PubMed Abstract | Publisher Full Text

41. Roberts M, Russell LB, Paltiel AD, *et al.*: **Conceptualizing a model: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--2.** *Value Health.* 2012; **15**(6): 804–11.
PubMed Abstract | Publisher Full Text | Free Full Text

42. Eddy DM, Hollingworth W, Caro JJ, *et al.*: **Model transparency and validation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--7.** *Value Health.* 2012; **15**(6): 843–50.
PubMed Abstract | Publisher Full Text

43. Ramos MC, Barton P, Jowett S, *et al.*: **A Systematic Review of Research Guidelines in Decision-Analytic Modeling.** *Value Health.* 2015; **18**(4): 512–29.
PubMed Abstract | Publisher Full Text

44. Caro JJ, Briggs AH, Siebert U, *et al.*: **Modeling good research practices--overview: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--1.** *Value Health.* 2012; **15**(6): 796–803.
PubMed Abstract | Publisher Full Text

45. Mandelblatt JS, Cronin KA, Bailey S, *et al.*: **Effects of mammography screening under different screening schedules: model estimates of potential benefits and harms.** *Ann Intern Med.* 2009; **151**(10): 738–47.
PubMed Abstract | Publisher Full Text | Free Full Text

46. Althaus CL, Turner KM, Schmid BV, *et al.*: **Transmission of *Chlamydia trachomatis* through sexual partnerships: a comparison between three individual-based models and empirical data.** *J R Soc Interface.* 2012; **9**(66): 136–46.
PubMed Abstract | Publisher Full Text | Free Full Text

47. Hontelez JA, Lurie MN, Bärnighausen T, *et al.*: **Elimination of HIV in South Africa through expanded access to antiretroviral therapy: a model comparison study.** *PLoS Med.* 2013; **10**(10): e1001534.
PubMed Abstract | Publisher Full Text | Free Full Text

48. Law MG, Prestage G, Grulich A, *et al.*: **Modelling the effect of combination**

antiretroviral treatments on HIV incidence. *AIDS.* 2001; **15**(10): 1287–94.
PubMed Abstract | Publisher Full Text

49. Atkins D, Best D, Briss PA, *et al.*: **Grading quality of evidence and strength of recommendations.** *BMJ.* 2004; **328**(7454): 1490.
PubMed Abstract | Publisher Full Text | Free Full Text

50. Andrews JC, Schünemann HJ, Oxman AD, *et al.*: **GRADE guidelines: 15. Going from evidence to recommendation-determinants of a recommendation's direction and strength.** *J Clin Epidemiol.* 2013; **66**(7): 726–35.
PubMed Abstract | Publisher Full Text

51. Higgins JB, Altman DG, Gøtzsche PC, *et al.*: **The Cochrane Collaboration's tool for assessing risk of bias in randomised trials.** *BMJ.* 2011; **343**: d5928.
PubMed Abstract | Publisher Full Text | Free Full Text

52. Sterne JA, Hernán MA, Reeves BC, *et al.*: **ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions.** *BMJ.* 2016; **355**: i4919.
PubMed Abstract | Publisher Full Text | Free Full Text

53. Guyatt GH, Oxman AD, Vist G, *et al.*: **GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias).** *J Clin Epidemiol.* 2011; **64**(4): 407–15.
PubMed Abstract | Publisher Full Text

54. Aronson N, Grant MD: **Tools for health care decision making: observational studies, modeling studies, and network meta-analyses.** *Value Health.* 2014; **17**(2): 141–2.
PubMed Abstract | Publisher Full Text

Open Peer Review

Current Referee Status: ✔ ✔

Version 2

Referee Report 27 February 2018

doi:10.5256/f1000research.15275.r31178

✔ **Joseph W Hogan**
Department of Biostatistics and Center for Statistical Sciences, Brown University School of Public Health, Providence, RI, USA

Thanks for the detailed replies; the revision addresses all of my comments.

Competing Interests: No competing interests were disclosed.

Referee Expertise: Biostatistics

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Referee Report 15 November 2017

doi:10.5256/f1000research.13392.r25673

? **Joseph W Hogan**
Department of Biostatistics and Center for Statistical Sciences, Brown University School of Public Health, Providence, RI, USA

The authors propose to incorporate findings from mathematical modeling studies into the development of WHO guidelines and other processes related to evaluating and developing public health policies. They argue that evidence from modeling studies should be included in
the Grading of Recommendations Assessment, Development and Evaluation (GRADE) process, and make specific recommendations to that effect. The authors argue that model credibility is more appropriate than risk of bias for evaluating strength of evidence generated by modeling studies. The paper is based on discussions and findings from a meeting of modeling experts in Geneva in 2016; the authors were also participants in the meeting.

The paper lays out a structured argument for incorporating modeling studies into the evidence base, particularly for formulating WHO recommendations related to treatment of HIV. The authors start by providing a review of how models are used in various fields, with suggestions about how they can inform

guideline development. They address the question of what constitutes a modeling study. A comprehensive accounting of published literature on assessment of models is provided. Finally, they give recommendations for how models can be evaluated using the GRADE approach, with specific conclusions about important issues such as when modeling studies should be used as part of the evidence base and how their credibility should be assessed.

Given the sweeping variety of models used in published studies about HIV treatments, policies and interventions, the authors are to be applauded for putting forward a framework for having this conversation. It will promote broader understanding of how models work and how they can be most optimally used for informing treatment guidelines.

At the crux of their argument is the claim that evidence generated from models should be judged in terms of model credibility rather than on risk of bias. This argument raises several important issues. First, what constitutes a mathematical model? If a model is to be evaluated on its credibility, we need a definition to work from. Second, what kinds of output from models should be considered evidence, and how should the quality of that evidence be judged and ultimately weighed against or combined with evidence generated from randomized trials and observational cohort data?

What is a mathematical model?

According to the authors, mathematical modeling is “a mathematical framework representing variables and their interrelationships to describe observed phenomena or predict future events.”

On its face this is surely true, but for the purpose of understanding whether and how models should add to the evidence base, it’s too broad. This definition covers a vast assortment of mathematical models, ranging from validated descriptions of natural phenomena (where the mathematical relationships are known and directly observable) to representations of progression from HIV infection to death (comprising known and unknown mathematical components, many of which cannot be directly observed).

Consider three examples for illustration:

1. The mathematical representation of radioactive decay is known and can be written down explicitly. The model enables accurate and replicable predictions of future observations.
2. The mathematical model for absorption of a specific drug is typically not known, but empirical studies have shown that it is possible to approximate the systematic variation using nonlinear equations. These models incorporate known information about physiology and properties of a specific drug, but are necessarily oversimplified representations of drug absorption because there are unobservable characteristics of individuals that affect absorption. The models can be used to make reliable predictions on average, but require unexplained variation to be reflected in terms of prediction intervals.
3. Now consider a model of the population dynamics of HIV infection and disease progression. This process also follows a mathematical model, but the model itself is highly complex. Unlike radioactive decay or rate of drug absorption, the mathematical representations of several components of the underlying processes are essentially unknown. Moreover, much of the data needed to inform the models are either unobserved (e.g. timing of HIV infection) or only sparsely observed (e.g. individual-level viral load).

All of these are mathematical models, but definitions must distinguish between them. Otherwise there is an implied equivalence that lends more credibility than is deserved to models that are heavily reliant on

unverified assumptions about the mathematical structure underlying the dynamic system being modeled. A more systematic classification of model types would therefore be helpful.

While the authors' definition of mathematical model is overly broad, the definition of statistical model, used to contrast with mathematical models, is too narrow. A statistical model is used to characterize sources of variation in observed data. It is based on a probabilistic representation of the data generating mechanism, which is itself a mathematical model. Theory and methods of statistical inference provides a rigorous and transparent set of techniques for parameter estimation, prediction of future outcomes, extrapolation (e.g. for causal inference), and uncertainty quantification. The last of these, uncertainty quantification, is a critical and frequently missing component of predictions based on mathematical models.

For the purposes of generating evidence for WHO recommendations, the main difference between a mathematical model and a statistical model is that mathematical models tend to have broader scope and incorporate higher dimensions of complexity, but rely more heavily on assumptions about underlying mathematical structure than on individual-level data. Statistical models tend to have less mathematical complexity and more narrow scope, and are typically fitted to a single (possibly large) set of observed individual-level data drawn from the target population(s) of interest. A mathematical structure underlies both statistical and mathematical models, and both can be used for prediction of future outcomes and for causal policy comparisons.

Should models be judged on 'risk of bias'?

The authors propose that evidence generated from mathematical models should be weighted more heavily toward model credibility than risk of bias.

Many mathematical models are over-parameterized relative to the amount of data used to fit them; hence multiple configurations of parameter values can be lead to very similar predictions. Mathematical models are typically calibrated to observed population-level data (e.g. annual HIV incidence rate for the target population), but the formal rules for doing this seem to vary across application.

For many consumers of model-based outputs, this is a significant methodologic concern that goes directly to the question of credibility. If multiple model configurations can generate similar predictions, which configuration is the most credible one? It seems reasonable that model-based outcomes such as 10-year predictions of HIV incidence need to be evaluated on their own terms. If coupled with a formal process for back-checking or recalibrating existing models this would surely add value, and would possibly strengthen model identifiability (i.e., provide evidence in favor of one set of model parameters over another).

A more general justification for incorporating risk of bias into model evaluation can be found in Coveney et al [1] (page 4), who provide a general rubric for assessing quality of scientific evidence in the age of big data, emphasizing ‘acceptance of the theory based on concordance between the predictions and the measurements.’ Model calibrations at the time of model fitting partially fulfill this objective, but post-hoc evaluation of model predictions must play an important role in establishing credibility.

The process of combining and comparing models is highly innovative and likely to have a positive impact on whether the results will be well received. This kind of cooperation and collaboration, exemplified recently by the Modelling Consortium, is perhaps unique to the mathematical modeling community. Evidence generated by these kinds of activities can form an important part of the evidence base.

Summary

The authors have provided a thorough case for including results from mathematical modeling into the formal evidence base used for making health recommendations, especially as they relate to HIV. The paper is based on findings from a recent conference and a comprehensive survey of extant literature.

The main critiques are that the definition of mathematical model is far too broad, and that bias (or risk of bias) needs to be incorporated into the evaluation criteria. Formal methods for uncertainty quantification are critical as well.

Mathematical models are prevalent and influential in the HIV literature; hence a discussion about whether and how to place their findings in the broader evidence base is needed and welcome. This paper provides a necessary starting point.

References

1. Coveney PV, Dougherty ER, Highfield RR: Big data need big theory too.*Philos Trans A Math Phys Eng Sci*. 2016; **374** (2080). PubMed Abstract | Publisher Full Text

Is the topic of the opinion article discussed accurately in the context of the current literature?
Yes

Are all factual statements correct and adequately supported by citations?
Partly

Are arguments sufficiently supported by evidence from the published literature?
Yes

Are the conclusions drawn balanced and justified on the basis of the presented arguments?
Yes

Competing Interests: No competing interests were disclosed.

Referee Expertise: Biostatistics

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 13 Feb 2018

Susan Norris, World Health Organization, Switzerland

Joseph W Hogan, Department of Biostatistics and Center for Statistical Sciences, Brown University School of Public Health, Providence, RI, USA

Approved with Reservations

Authors' response: Thank you very much for reviewing our paper and making such thoughtful

comments. We address your reservations one by one below. We have made changes in the text via tracked changes.

The authors propose to incorporate findings from mathematical modeling studies into the development of WHO guidelines and other processes related to evaluating and developing public health policies. They argue that evidence from modeling studies should be included in the Grading of Recommendations Assessment, Development and Evaluation (GRADE) process, and make specific recommendations to that effect. The authors argue that model credibility is more appropriate than risk of bias for evaluating strength of evidence generated by modeling studies. The paper is based on discussions and findings from a meeting of modeling experts in Geneva in 2016; the authors were also participants in the meeting.

Authors' response: First, we would like to stress that our article is an opinion piece, and does not reflect an official position of the World Health Organization or any other body. In the introduction we write: "In this paper, which reflects the opinions of the authors… ". Also, we tried to keep the recommendations fairly general, rather than specific and prescriptive. For example, we refrained from recommending a specific instrument to assess the quality of modeling studies. We conclude by saying that "We look forward to discussing these recommendations with experts and stakeholders and to developing exact procedures and criteria for the assessment of modelling studies and their inclusion in the GRADE process."

The paper lays out a structured argument for incorporating modeling studies into the evidence base, particularly for formulating WHO recommendations related to treatment of HIV. The authors start by providing a review of how models are used in various fields, with suggestions about how they can inform guideline development. They address the question of what constitutes a modeling study. A comprehensive accounting of published literature on assessment of models is provided. Finally, they give recommendations for how models can be evaluated using the GRADE approach, with specific conclusions about important issues such as when modeling studies should be used as part of the evidence base and how their credibility should be assessed.

Authors' response: Thank you, this is a nice outline of the paper.

Given the sweeping variety of models used in published studies about HIV treatments, policies and interventions, the authors are to be applauded for putting forward a framework for having this conversation. It will promote broader understanding of how models work and how they can be most optimally used for informing treatment guidelines.

Authors' response: Thank you very much.

At the crux of their argument is the claim that evidence generated from models should be judged in terms of model credibility rather than on risk of bias. This argument raises several important issues. First, what constitutes a mathematical model? If a model is to be evaluated on its credibility, we need a definition to work from. Second, what kinds of output from models should be considered evidence, and how should the quality of that evidence be judged and ultimately weighed against or combined with evidence generated from randomized trials and observational cohort data?

Authors' response: We agree that these are central issues. Regarding the outputs from models that should be considered evidence, please note that we distinguish between *mathematical models* and *modelling studies*. The latter address well defined questions and outcomes, such as the impact of HIV testing and immediate antiretroviral therapy on HIV incidence or the impact of different screening strategies on the incidence of cervical cancer. In other words, the modeling outputs that constitute relevant evidence will depend on the question addressed in the modeling studies. In the revised version we write.

GRADE provides a well-defined framework for weighing evidence from randomized trials and observational studies, as discussed in the section on "Assessing mathematical modelling studies using the GRADE approach". Of note, randomized trials and observational studies are assessed separately. In general, guideline development groups will focus on randomized evidence if such evidence is available from several trial, and only consider observational studies in the absence of substantial randomized evidence. Similarly, evidence from mathematical modelling studies will be considered primarily if other studies cannot answer the question. Statistically combining evidence from different study types is not foreseen in GRADE, and beyond the scope of our article.

What is a mathematical model?

According to the authors, mathematical modeling is "a mathematical framework representing variables and their interrelationships to describe observed phenomena or predict future events."

On its face this is surely true, but for the purpose of understanding whether and how models should add to the evidence base, it's too broad. This definition covers a vast assortment of mathematical models, ranging from validated descriptions of natural phenomena (where the mathematical relationships are known and directly observable) to representations of progression from HIV infection to death (comprising known and unknown mathematical components, many of which cannot be directly observed).

Consider three examples for illustration:

1. The mathematical representation of radioactive decay is known and can be written down explicitly. The model enables accurate and replicable predictions of future observations.
2. The mathematical model for absorption of a specific drug is typically not known, but empirical studies have shown that it is possible to approximate the systematic variation using nonlinear equations. These models incorporate known information about physiology and properties of a specific drug, but are necessarily oversimplified representations of drug absorption because there are unobservable characteristics of individuals that affect absorption. The models can be used to make reliable predictions on average, but require unexplained variation to be reflected in terms of prediction intervals.
3. Now consider a model of the population dynamics of HIV infection and disease progression. This process also follows a mathematical model, but the model itself is highly complex. Unlike radioactive decay or rate of drug absorption, the mathematical representations of several components of the underlying processes are essentially unknown. Moreover, much of the data needed to inform the models are either unobserved (e.g. timing of HIV infection) or only sparsely observed (e.g. individual-level viral load).

All of these are mathematical models, but definitions must distinguish between them. Otherwise there is an implied equivalence that lends more credibility than is deserved to models that are heavily reliant on unverified assumptions about the mathematical structure underlying the dynamic system being modeled. A more systematic classification of model types would therefore be

helpful.
Authors' response: Thank you for these three examples. We agree that the definition by Pieter Eykhoff is fairly broad and covers all three categories. However, we feel it is clear from the title and text of our paper that we are primarily concerned with models of the second and third category, i.e. more complex mathematical models that are relevant to WHO guidelines. Within these categories, the level of abstraction and complexity of models and their credibility will of course vary (see also examples in Table 1).

We have now made explicit the distinction that we make between *mathematical models* and *mathematical modelling studies* at the beginning of the section, "What is a mathematical modelling study?" (page 4):

A broad definition of a mathematical model is a "mathematical framework representing variables and their interrelationships to describe observed phenomena or predict future events".[5] We make a distinction between a mathematical model and mathematical modelling studies, which we define as studies that address defined research questions using mathematical models with a considerable degree of complexity and abstraction.

At the Geneva workshop participants discussed different types of mathematical models in detail, based on a presentation by one of the authors (CA) on "The anatomy of mathematical modelling studies". The workshop report and slides and be found at http://apps.who.int/iris/bitstream/10665/258987/1/WHO-HIS-IER-REK-2017.2-eng.pdf. Unfortunately, the link to the report and slides was incorrect in the F1000research paper. CA discussed model dichotomies, based on the book by Ben Bolker (Ecological Models and Data in R, 2008, Princeton University Press), and illustrated these using case studies from the Ebola crisis – see copy of one of his slides at the end of this response. In the discussion, workshop participants argued that guideline groups will often not be able to differentiate between different model dichotomies, and that this is not essential: guideline groups "just have to know what information models can provide and what value can be placed in that information." However, we recommend that experts in mathematical modelling should support guideline groups (see recommendation 5 on page 9).

We agree with the referee that guideline developers should carefully assess the credibility of models, and that models that "are heavily reliant on unverified assumptions about the mathematical structure underlying the dynamic system" are not credible. Our review of the methodological literature (see Table 2 in the paper) showed that the published frameworks of good modelling practice consistently emphasize the importance of the rationale for the model structure, the structural assumptions and uncertainty, the model transparency and validation etc. See also our recommendations 4, 5 and 6 on p 9.

We added a more explicit reference and the correct link to the Workshop report, (page 3, last line):

A detailed workshop report is available from WHO[4].

We also expanded the section, "What is a mathematical modelling study" to clarify our view on the need for classifying mathematical modelling studies (page 4, last paragraph):

Workshop participants discussed whether it might be helpful for guideline groups to classify mathematical models in terms of their scope (for example descriptive versus

predictive) or technical approach (for example static versus dynamic)[8]. Discussants argued that a good understanding of what information models can provide and what level of confidence can be placed in that information was more important than a taxonomy of models[4].

While the authors' definition of mathematical model is overly broad, the definition of statistical model, used to contrast with mathematical models, is too narrow. A statistical model is used to characterize sources of variation in observed data. It is based on a probabilistic representation of the data generating mechanism, which is itself a mathematical model. Theory and methods of statistical inference provides a rigorous and transparent set of techniques for parameter estimation, prediction of future outcomes, extrapolation (e.g. for causal inference), and uncertainty quantification. The last of these, uncertainty quantification, is a critical and frequently missing component of predictions based on mathematical models.

Authors' response: We agree with the reviewer's comment about assessing uncertainty in mathematical models and state this explicitly (page 5, paragraph 2).

For the purposes of generating evidence for WHO recommendations, the main difference between a mathematical model and a statistical model is that mathematical models tend to have broader scope and incorporate higher dimensions of complexity, but rely more heavily on assumptions about underlying mathematical structure than on individual-level data. Statistical models tend to have less mathematical complexity and more narrow scope, and are typically fitted to a single (possibly large) set of observed individual-level data drawn from the target population(s) of interest. A mathematical structure underlies both statistical and mathematical models, and both can be used for prediction of future outcomes and for causal policy comparisons.

Authors' response: We are grateful to the referee for this insightful and well-phrased comment about the relevance of the terms 'statistical modelling' and 'mathematical modelling' to WHO guidelines. We have taken the liberty of paraphrasing the comment to revise this section (page 4) as follows:

Mathematical modelling studies may address complex situations and tend to rely more heavily on assumptions about underlying mathematical structure than on individual-level data. Examples include investigating the potential of HIV testing with immediate antiretroviral therapy to reduce HIV transmission[6], or the likely impact of different screening practices on the incidence of cervical cancer[7].

Statistical modelling is typically concerned with characterizing sources of variation and associations between variables in observed individual-level data drawn from a target population of interest. Statistical models tend to be narrower in scope than mathematical models. Both statistical and mathematical models can be used to predict future outcomes and to compare different policies. The results from statistical analyses of empirical data often inform mathematical models. Mathematical modelling studies also increasingly integrate statistical models to relate the model output to data.

Should models be judged on 'risk of bias'?

The authors propose that evidence generated from mathematical models should be weighted more heavily toward model credibility than risk of bias.

Authors' response: Yes, we believe that the concept of model credibility is more useful than the more narrow concept of risk of bias (RoB). However, we think there is a mis-understanding here: the assessment RoB of specific studies also has a role.

The RoB concept is widely used in the context of randomized controlled trials and observational studies that aim to make causal inference, and dedicated "RoB tools" have been developed to assess the risk of bias of studies included in systematic reviews (see references 1,2 below and www.riskofbias.info). These tools are based on relatively few well-defined biases. In the case of randomized trials they include selection bias, performance bias, detection bias, attrition bias and reporting bias (1).

In the context of mathematical modelling studies, the risk of bias of empirical studies contributing parameter estimates is important and should be considered, for example in sensitivity analyses. On the other hand, many other and additional aspects are important when assessing the trustworthiness or credibility of mathematical models. These aspects are listed in Table 2, based on a review of published frameworks developed to assess good modelling practice. Please note that we use the term credibility as applied by the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) to assessment of studies for decision making (3).

These frameworks include assessments of the quality of the data used to parameterize a model. For example, Bennett and Manuel (4) and Philips et al (5) include several questions to that effect:

- Where choices have been made between data sources, are these justified appropriately?
- Where data from different sources are pooled, is this done in a way that the uncertainty relating to their precision and possible heterogeneity is adequately reflected?
- Has the quality of the data been assessed appropriately?

The questionnaire proposed by Caro et al (6) asks

- Are the data used in populating the model suitable for your decision problem?
- All things considered, do you agree with the values used for the inputs?

Similarly, the framework of Ramos and colleagues (7) includes the following questions:

- Have transition probabilities and intervention effects been derived from representative data sources for the decision problem?
- Have parameters relating to the effectiveness of interventions derived from observational studies been controlled for confounding?

We have clarified our position and the role of RoB assessments as follows on page 8:

The concept of bias relates to results or inferences from empirical studies, including randomized controlled trials and observational studies[38,39] and is too narrow in the context of assessing mathematical modelling studies.[40] "Credibility", a term used by ISPOR,[41] may therefore be more appropriate for modelling studies than "risk of bias". The assessment of the credibility of a model is informed by a comprehensive quality framework and should cover the conceptualization of the problem, model structure, the input data and their risk of bias, different dimensions of uncertainty, as well as transparency and validation (Table 2).

1. Higgins JP, Altman DG, Gotzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011; 343:d5928.
2. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016; 355:i4919.
3. Aronson N, Grant MD. Tools for health care decision making: observational studies, modeling studies, and network meta-analyses. Value Health 2014; 17:141–142.
4. Bennett C, Manuel DG: Reporting guidelines for modelling studies. BMC Med Res Methodol. 2012; 12: 168.
5. Philips Z, Bojke L, Sculpher M, et al.: Good practice guidelines for decision- analytic modelling in health technology assessment: a review and consolidation of quality assessment. Pharmacoeconomics. 2006; 24(4): 355–71.
6. Jaime Caro J, Eddy DM, Kan H, et al.: Questionnaire to assess relevance and credibility of modeling studies for informing health care decision making: An ISPOR-AMCP-NPC good practice task force report. Value Health. 2014; 17(2): 174–82.
7. Ramos MC, Barton P, Jowett S, et al.: A Systematic Review of Research Guidelines in Decision-Analytic Modeling. Value Health. 2015; 18(4): 512–29.

Many mathematical models are over-parameterized relative to the amount of data used to fit them; hence multiple configurations of parameter values can be lead to very similar predictions. Mathematical models are typically calibrated to observed population-level data (e.g. annual HIV incidence rate for the target population), but the formal rules for doing this seem to vary across application.

Authors' response: We agree – model concept, structure and parsimony are important elements when evaluating the credibility of mathematical models. Validation and predictive validity are also very important – again see Table 2.
For many consumers of model-based outputs, this is a significant methodologic concern that goes directly to the question of credibility. If multiple model configurations can generate similar predictions, which configuration is the most credible one? It seems reasonable that model-based outcomes such as 10-year predictions of HIV incidence need to be evaluated on their own terms. If coupled with a formal process for back-checking or recalibrating existing models this would surely add value, and would possibly strengthen model identifiability (i.e., provide evidence in favor of one set of model parameters over another).

Authors' response: We agree and, again, believe that these issues are covered by the frameworks we present in Table 2.

A more general justification for incorporating risk of bias into model evaluation can be found in Coveney et al [1] (page 4), who provide a general rubric for assessing quality of scientific evidence in the age of big data, emphasizing 'acceptance of the theory based on concordance between the predictions and the measurements.' Model calibrations at the time of model fitting partially fulfill this objective, but post-hoc evaluation of model predictions must play an important role in establishing credibility.

Authors' response: The timely piece by Coveney and Dougherty is really a critique of "'blind' big data projects" and a plea for "the elucidation of the multiscale and stochastic processes controlling the behaviour of complex systems, including those of life, medicine and healthcare." We could not agree more and argue that insights from the latter (mathematical models) should inform the

development of WHO guidelines.

The process of combining and comparing models is highly innovative and likely to have a positive impact on whether the results will be well received. This kind of cooperation and collaboration, exemplified recently by the Modelling Consortium, is perhaps unique to the mathematical modeling community. Evidence generated by these kinds of activities can form an important part of the evidence base.

Authors' response: We completely agree and have stressed this point in our paper.

Summary

The authors have provided a thorough case for including results from mathematical modeling into the formal evidence base used for making health recommendations, especially as they relate to HIV. The paper is based on findings from a recent conference and a comprehensive survey of extant literature.

The main critiques are that the definition of mathematical model is far too broad, and that bias (or risk of bias) needs to be incorporated into the evaluation criteria. Formal methods for uncertainty quantification are critical as well.

Authors' response: Thank you again for reviewing our paper. See our responses above. We hope that based on our responses and the changes made in the manuscript you will be able to approve our contribution.

Mathematical models are prevalent and influential in the HIV literature; hence a discussion about whether and how to place their findings in the broader evidence base is needed and welcome. This paper provides a necessary starting point.

Competing Interests: No competing interests were disclosed.

Referee Report 20 September 2017

doi:10.5256/f1000research.13392.r25463

Wilma A. Stolk
Erasmus MC, Department of Public Health, University Medical Center Rotterdam, Rotterdam, Netherlands

In this opinion article, the authors discuss when and how to incorporate the results of modelling studies into WHO guidelines, by addressing three questions: (1) When is it appropriate to consider modelling studies as part of the evidence that supports a guideline? (2) How should the quality and risk of bias in mathematical modelling studies be assessed? (3) How can the GRADE approach be adapted to assess the certainty of a body of evidence that includes the results of modelling and to formulate recommendations? Based on findings from a web-based expert survey, a rapid literature review to identify criteria for assessing the "quality" of mathematical modelling studies, and on discussions and presentations at a workshop on the topic that was held April 2016 in Geneva, the authors conclude that modelling studies should indeed routinely be considered in the process of developing WHO guidelines,

particularly in the evaluation of public health programmes, long-term effectiveness or comparative effectiveness. As for other types of evidence taken into consideration, there should be a systematic and transparent approach to identifying existing models that may be relevant and the quality and credibility of models should be systematically assessed. Relatively few adaptations are needed in the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach to rate the certainty of a body of evidence and to produce information that is used by guideline panels to formulate recommendations, based on the balance of benefits and harms and other considerations.

MINOR COMMENTS:

1. Recommendation 4 is “No single ‘one-size-fits-all’ approach is appropriate to assess the quality of modelling studies. Existing frameworks and checklists may be adapted to a set of modelling studies by adding or omitting questions. In some situations, the approach will need to be developed *de novo*.” I’d prefer to turn it around: based on existing frameworks and checklists, generic criteria can be developed to assess the quality of modelling studies, although – depending on the situation –questions may have to be added or omitted. I am not convinced that in some situations a completely new approach is needed, and this would also not be advisable. The authors should either delete the last statement, or explain under which circumstances such a new approach is needed, ideally illustrated with an example.

2. Recommendation 8 is “The certainty of the evidence for modelling studies should be assessed and presented separately in summaries of the evidence (GRADE evidence profiles), and classified as high, moderate, low, or very low certainty.” In the text, the authors state that RCTs start as high certainty and observational studies as low certainty, although this certainty score may be up- or down-rated based on detailed assessment of five dimensions. Is it possible to give an indication of where modelling studies would start, with a justification? If not, can the authors describe factors to be considered when determining the start class?

3. The questionnaire of the online survey on the use of mathematical modelling in guidelines for public health decision making is included as Figure S1, which combines a series of screen shots. The quality of this figure is poor and I recommend to include the questionnaire as a text document.

Is the topic of the opinion article discussed accurately in the context of the current literature?
Yes

Are all factual statements correct and adequately supported by citations?
Yes

Are arguments sufficiently supported by evidence from the published literature?
Yes

Are the conclusions drawn balanced and justified on the basis of the presented arguments?
Yes

Competing Interests: No competing interests were disclosed.

Referee Expertise: Epidemiology / mathematical modelling, with focus on neglected tropical diseases

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 13 Feb 2018

Susan Norris, World Health Organization, Switzerland

Reviewer 1: Wilma A. Stolk, Erasmus MC, Department of Public Health, University Medical Center Rotterdam, Rotterdam, Netherlands

Approved

Authors' response: Thank you for reviewing and approving our paper. See below our responses to your comments. We have made changes in the text via tracked changes.

In this opinion article, the authors discuss when and how to incorporate the results of modelling studies into WHO guidelines, by addressing three questions: (1) When is it appropriate to consider modelling studies as part of the evidence that supports a guideline? (2) How should the quality and risk of bias in mathematical modelling studies be assessed? (3) How can the GRADE approach be adapted to assess the certainty of a body of evidence that includes the results of modelling and to formulate recommendations? Based on findings from a web-based expert survey, a rapid literature review to identify criteria for assessing the "quality" of mathematical modelling studies, and on discussions and presentations at a workshop on the topic that was held April 2016 in Geneva, the authors conclude that modelling studies should indeed routinely be considered in the process of developing WHO guidelines, particularly in the evaluation of public health programmes, long-term effectiveness or comparative effectiveness. As for other types of evidence taken into consideration, there should be a systematic and transparent approach to identifying existing models that may be relevant and the quality and credibility of models should be systematically assessed. Relatively few adaptations are needed in the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach to rate the certainty of a body of evidence and to produce information that is used by guideline panels to formulate recommendations, based on the balance of benefits and harms and other considerations.

Authors' response: Thank you. This is a nice summary of our paper.

MINOR COMMENTS:

1. Recommendation 4 is "No single 'one-size-fits-all' approach is appropriate to assess the quality of modelling studies. Existing frameworks and checklists may be adapted to a set of modelling studies by adding or omitting questions. In some situations, the approach will need to be developed *de novo*." I'd prefer to turn it around: based on existing frameworks and checklists, generic criteria can be developed to assess the quality of modelling studies, although – depending on the situation –questions may have to be added or omitted. I am not convinced that in some situations a completely new approach is needed, and this would also not be advisable. The authors should either delete the last statement, or explain under which circumstances such a new approach is needed, ideally illustrated with an example.

Authors' response: Thank you. We agree and have deleted the last statement on page 9.

1. Recommendation 8 is "The certainty of the evidence for modelling studies should be assessed and presented separately in summaries of the evidence (GRADE evidence profiles), and classified as high, moderate, low, or very low certainty." In the text, the authors state that RCTs start as high certainty and observational studies as low certainty, although this certainty score may be up- or down-rated based on detailed assessment of five

dimensions. Is it possible to give an indication of where modelling studies would start, with a justification? If not, can the authors describe factors to be considered when determining the start class?

Authors' response: Thank you, we have addressed this issue as follows on page 9:
"We propose that within the GRADE system, modelling studies start at low certainty, and it is then possible to increase or decrease the certainty of modelling studies based on a set of criteria. The development of these criteria was beyond the scope of this article; a GRADE working group is addressing this issue (http://www.gradeworkinggroup.org/)."

1. The questionnaire of the online survey on the use of mathematical modelling in guidelines for public health decision making is included as Figure S1, which combines a series of screen shots. The quality of this figure is poor and I recommend to include the questionnaire as a text document.

Authors' response: Thank you. There is no text document for the survey but we have enlarged the screen shots to increase their readability (pages 22-27).

Competing Interests: No competing interests were disclosed.